Terminology Issues in User Access to Web-based Medical Information

Alexa T. McCray, Russell F. Loane, Allen C. Browne, Anantha K. Bangalore
National Library of Medicine
Bethesda, Maryland

We conducted a study of user queries to the National Library of Medicine Web site over a three month period. Our purpose was to study the nature and scope of these queries in order to understand how to improve users' access to the information they are seeking on our site. The results show that the queries are primarily medical in content (94%), with only a small percentage (5.5%) relating to library services, and with a very small percentage (.5%) not being medically relevant at all. We characterize the data set, and conclude with a discussion of our plans to develop a UMLS-based terminology server to assist NLM Web users.

INTRODUCTION

Understanding the behavior of users as they access Web-based information systems can be critical to the success of the system and can enhance its ability to deliver the information it contains. It has become more and more difficult for Web users to find relevant information in the large information space they are now forced to navigate. Users become frustrated when they do not find what they need quickly, and they often do not know how to improve their search strategies when they have retrieved too much information, no information, or information that is not directly relevant to their questions [1,2].

The U.S. National Library of Medicine (NLM) maintains and continuously updates the information resources it makes available from its Web site (http://www.nlm.nih.gov/). NLM's Web resources range from general information about the NLM, such as library hours, staff directories, and job and training opportunities, to information about special programs such as the Visible Human project, Telemedicine projects, and the Unified Medical Language System (UMLS), to publications, such as newsletters, fact sheets, and specialized bibliographies, to, importantly, its many databases, including MEDLINE, GenBank, AIDSTRIALS, and TOXNET. In Tannery and Wessel's [3] terminology, NLM's site is a Level III site, since it contains not only basic information about the library, but also original data and locally generated databases. Figure 1 below shows the current NLM home page.

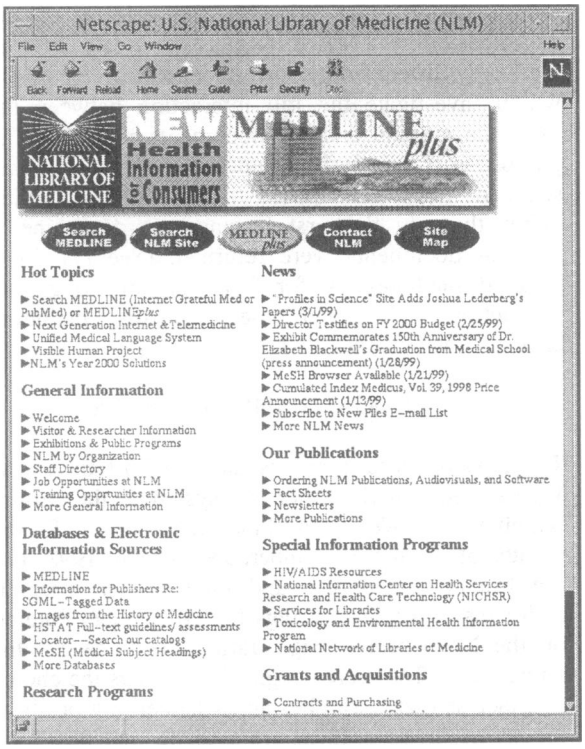

Figure 1. NLM Home Page
http://www.nlm.nih.gov/

In order to search individual NLM databases, users must first choose the link to the database of interest from the home page, e.g., MEDLINE, and then the interface to that database is presented and searching may proceed. The link on the NLM home page marked "Search NLM site" allows users to search general information, fact sheets, and bibliographies, but it does not search the databases directly, since each of the databases has its own specialized interface and backend engine. However, the home page does provide navigation help as well as information about the content of the databases that are available.

In late October 1998, NLM announced its new resource MEDLINEplus, intended primarily for consumers of health information. The site organizes information according to common health topics, such as AIDS, asthma, cancers, heart attacks, and obesity,

1091—8280/99/$5.00

and it provides links to medical dictionaries, clearinghouses of consumer health literature, and directories for finding health care professionals and facilities.

In an effort to better understand what users are interested in finding on the NLM site, we undertook a study of the NLM Web query log files over a three month period, just prior to the release of the consumer-oriented MEDLINEplus site in late October. Our study differs from some other studies, e.g.,[4-6], which investigate the full interaction between the user, the system, and the documents returned. The purpose of this analysis is to determine the nature and scope of the queries posed to the NLM Web site, rather than, in the first instance, to determine if relevant documents were returned. (See [7] for a detailed methodology for fully instrumenting user interactions with Web-based systems, including critical privacy concerns.)

MATERIALS AND METHODS

The Apache Web server is used to host the NLM Web services, and the ht://Dig system is the current search engine. We analyzed the access log file for the months of August, September, and October 1998. The log file contained date and time stamps, user IP addresses, and search strings for every query entered on the NLM search page during the three months under study. The search engine gives users the choice of exact or fuzzy searches and whether any or all of the terms in their query were to be matched (i.e., Boolean OR or AND). These options were not considered further in our analysis, since, as noted by others (e.g., [8]), users of Web search engines tend not to use these options with any degree of consistency.

First, we excluded all queries with blank or non-text search strings, as well as any queries submitted from an NLM IP address. This resulted in a total of 336,990 queries made to the search system (using the "Search NLM Site" button) during the three month period. We loaded the data into an ORACLE database for further analysis and manipulation. Each query string was normalized using the UMLS lexical programs [9], which ignore case, punctuation, word order, and singular/plural variation. Multiple queries with the same normalized string and IP address were merged and counted as one unique query. This approach eliminated duplication arising from the same queries issued by the same user on different days and also suppressed duplication arising from users stepping through many pages of search results. The process resulted in a set of 225,164 unique queries.

We then mapped the uniqued queries to UMLS concepts through calls to the UMLS Knowledge Source Server [10]. In some cases, queries map to multiple UMLS concepts. For example, the user query *dressing* maps to several concepts in the UMLS, including *clothing assistance* as a health care activity, and *dressing of skin or wound* as a therapeutic procedure. When N concepts matched, we applied a 1/N weighting factor when calculating the frequency of occurrence of the concept in the queries. This weighting represents the uncertainty about which concept was actually intended by the user. For the queries that mapped to UMLS concepts, we captured the semantic types of those concepts and grouped them into fourteen higher-level semantic groups. These groups are shown in the results section below.

Second, the normalization process reduced the 225,164 queries to 128,640 unique normalized strings. We mapped these unique strings to the UMLS using the same process described above. In addition, we checked all strings that did not map to a UMLS concept against the Metathesaurus normalized word index to see if the strings were found as constituents of at least one Metathesaurus concept.

Third, we carried out a number of additional analyses that would allow us to give an overall characterization of the types of queries that users submitted. We reviewed all strings that appeared with a frequency of greater than one to determine which were appropriate to the NLM home page and which were general medical questions. Word counts of the queries in the entire data set were made and compared to the terms found in the Metathesaurus. We studied a subset of the queries to determine whether users iterated with the system in a query session, and, if so, what the nature of that iteration was. After scanning the data, we arbitrarily chose a query session to continue until a delay of thirty or more minutes elapsed between queries.

RESULTS

Some 91,944 (41%) of the 225,164 unique, normalized queries mapped successfully to the Metathesaurus, while 133,220 (59%) did not map directly. For those that did map to concepts, the largest number were diseases and other types of disorders, with drugs being the second most frequent semantic type category. The results of the semantic type distribution are shown below in Figure 2.

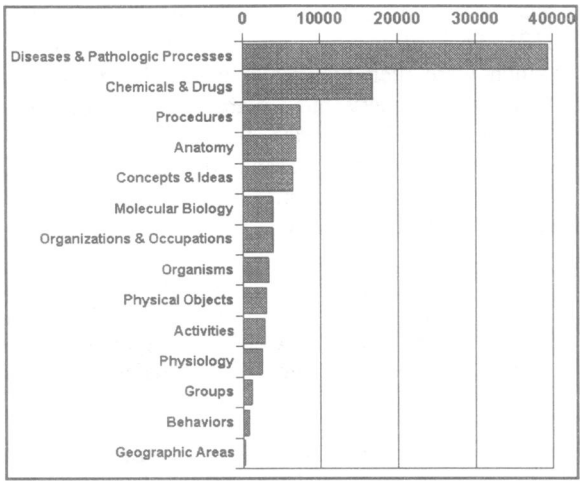

Figure 2. Distribution of semantic type groupings for 91,944 queries

Figure 2 illustrates that, for those queries that mapped to UMLS concepts, diseases and their treatments (drugs and procedures) appeared to be of greatest interest. Some of the anatomy terms may also have referred, in abbreviated form, to diseases. Thus, for example, *heart, prostate*, and *gall bladder* might actually have been requests for information about heart disease, prostate cancer and gall bladder problems.

Mapping the 128,640 unique normalized strings to the Metathesaurus resulted in a match of 20,754 (16%) to Metathesaurus concepts and 107,886 (84%) which did not match. In order to assess the possibility that the non-matches were found in some form in the Metathesaurus, we used the normalized word index to look for the 9,986 strings which did not match Metathesaurus concepts directly and which appeared with a frequency > 1. Some 30% of the non-matching search strings were found in their entirety as constituents of Metathesaurus concepts. A few examples illustrate. The terms labeled "a" in each pair are the search strings and the terms labeled "b" are the Metathesaurus concepts of which the strings are constituents. In many cases the string was found in multiple concepts, but for simplicity, a single representative is shown below.

1a) *head and neck melanoma*
1b) *malignant melanoma of head and neck*

2a) *intravenous techniques*
2b) *warming intravenous fluid technique*

3a) *dexterity test*
3b) *extremity testing for strength, dexterity or stamina*

We were interested to see whether the queries made to the NLM system were primarily medical in nature, whether they were related directly to the services provided by NLM, or whether they were not relevant to the site at all. We reviewed all 19,855 normalized strings of frequency > 1 and marked them according to the distinctions just noted. A surprisingly large number, 18,757 (94%), were medical terms. Some 1,098 (5.5%) related to NLM services, and only 95 (.5%) were not relevant to the site at all. The medical terms exhibited the range of categories already seen above, including a large percentage of diseases and their treatments.

The terms related to NLM services were queries about journals (e.g., *MEDLINE list journals, Archives of Diseases in Childhood, psychiatric journals*), queries about specific NLM resources (e.g., *NLM Technical Bulletin, medlars II file descriptions, bioethics thesaurus*), queries about general library services (e.g., *library catalogue, document delivery, library holdings*), other queries related to program areas within the NLM, such as grants (e.g., *information system grant*), the history of medicine division (e.g.,*Islamic culture and the medical arts*), and the library operations division (e.g., *medlars management*), and, finally, some number of the queries were the names of people on the NLM staff.

Some examples of the very small number of queries not obviously relevant to the site are *time management, car seat,* and *smile*, together with just a handful of taboo words. A list of the top twenty terms, in descending order of frequency, is given in the appendix.

Some interesting patterns emerged in the analysis of these terms. First, many terms were misspelled, sometimes quite badly. This is consistent with what other investigators have found (e.g., [1]). For example, the term *prostate cancer* was often misspelled as *prostrate cancer*; *leukemia* was misspelled as *lukemia, leukimia* and *luekemia*; *hepatitis* was variously misspelled as *hepetitis, hepitius, hepititus, and hepetitus.* One of the most frequently occurring queries, *fibromyalgia*, was misspelled as f*ybromyalgia, fibermyalgia, fibromalgia,* and *fibromyalsia*. Many other terms were misspelled, e.g., *farmacology, cartlidge, , psicology, ginecology, complimentary medicine*, and *tempeture.*

Quite often in the case of eponymic terms such as *Alzheimer's disease, Parkinson's disease*, and *Crohn's disease* only the name was used (e.g., *Alzheimer,* for *Alzheimer's disease*), rather than the full term. In the case of *Crohn's disease*, many variant spellings were used as well. All of the following appeared in the data: *Chron's Disease, Crone's disease, Chrone's Disease, crohn's desease, chromes disease, Krone's disease, chrones desease, Crohn's, crone, crones, crohnes, Chrohn's,* and *chrons*. Other types of abbreviatory terms or word fragments were also used, e.g., *cranio, divertic*, and *corti.*

There were many queries that expressed the same concept in a variety of ways. Queries about the visible human project were expressed, for example, as *virtual human, visual human, virtual body*, and even *invisible man.*

For all of the 128,640 unique normalized strings, we counted the number of words in each term in order to get a measure of the complexity of the queries. Figure 3 below shows the distribution of the number of words per term.

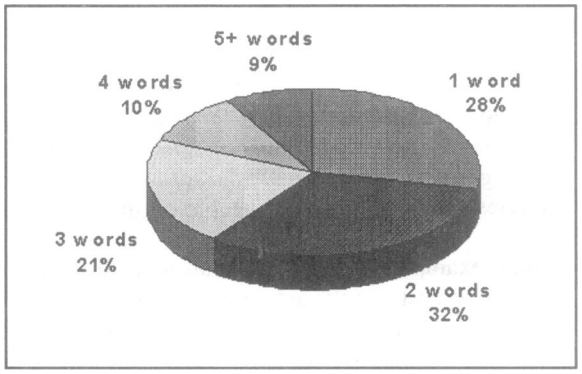

Figure 3. Distribution of number of words per term in 128,640 strings

The large majority of queries consist of one (28%), two (32%), or three (21%) words. We compared this with the strings in the 1999 Metathesaurus. The distribution is in sharp contrast to the query terms under study here. Some 29% of the Metathesaurus terms contain five or more words, while only 9% consist of a single word.. Many investigators have noted that the majority of terms submitted to Web systems are limited to one or two words (e.g., [6,11]). It is interesting to contrast this phenomenon with search queries as they existed when search specialists were involved. Some years ago, we did a study of over 150 search requests as part of our work in establishing an information retrieval test collection [12]. The search requests were expressed on forms which were then handed to a search specialist who translated the query into a search strategy and executed the search. The queries were quite detailed in comparison with the searches that are currently done by users of Web systems. Two examples of search requests are given below:

> *Please do literature search for any relationship between chloroquine and low blood pressure in people with pre-existing hypertension. Also possible interaction with diuretics to exaggerate hypotensive effect.*
>
> *Adaptation of physical environment in hospitals to care for Alzheimer patients (model Alzheimer units).*

In the above examples, it is possible to get a sense for the information need of the user, in a way that is rarely possible in today's search environment. It occurred to us that it might be interesting to study user sessions to see if we could infer something about the underlying information need. A sample session for one user over the period of about 35 minutes follows.

> *what is a chem.12 blood test and why is it done?*
> *Blood test called chem.12 or panel 12*
> *Blood test*
> *types of blood test for kidney damage*
> *types of blood test kidney damage*
> *types of blood test*
> *blood test*
> *chem.12, panel 12 blood test*
> *blood test for drugs and blood test for kidney function*
> *blood test for kidney damage*
> *types of blood test for kidney function*
> *types of blood test*
> *Blood test*

Each individual query was treated in isolation by the search engine, but it is clear that the user is concerned about a blood test for a type of kidney dysfunction. This session is interesting in a number of respects. The user starts with a rather complex query and then narrows it down to its essentials in hopes of finding some relevant information. Note that the user repeats blood test three times, perhaps in frustration. Pollock [1:section 5] suggests that "searching on the Internet is a process rather than an event", and that this should have implications for how future search engines are designed. She says that the idea is to "engineer this view of search as a process.

CONCLUSIONS

Our investigation of a large sample of queries made to the NLM home page reveals that users are primarily asking medical questions. That is, if they find the NLM Web site, they are asking questions that are appropriate to a medical site. They are not so much interested in the NLM as an institution, but rather as a rich repository of medical information.

The data show, however, that there is a mismatch in two respects between users' queries and the resources they are attempting to access. First, the queries are often very specific questions whose answers may or may not be contained in NLM's databases, and if they are, they may only be there indirectly, as would be the case with a MEDLINE search. Second, the terminology, while medical, is often not found directly in medical terminologies, such as those represented in the UMLS. The reasons for the mismatch vary. There are many misspellings, partial words, and abbreviatory forms. In addition, users of Web systems tend to use a small number of words in formulating their queries, and these may be insufficient to fully express the information need.

To address the first problem, NLM has established and continues to build its MEDLINEplus site. The resource is intended to make it easier for consumers to find health related information, whether in NLM databases or on other sites. Extensive navigation help is provided, together with specific search capabilities.

To address the second problem, we intend to explore the development of a terminology server whose goal it is to mediate between user terminology and terminology as it is reflected in a variety of medical information resources. The server is planned to be an application of the UMLS Knowledge Source Server, and it will need to address all of the issues noted in the analysis of the current data set, including misspellings, incomplete terms, missing synonymy, and terminology browsing capabilities.

REFERENCES

1. Pollock A, Hockley A. What's wrong with Internet searching. D-Lib Magazine, March 1997. http://www.dlib.org/dlib/march97/bt/03pollock.html.

2. Nielsen J. Search and you *may* find. http://www.sun.com/columns/jakob/, July 15, 1997.

3. Tannery NH, Wessel CB. Academic medical center libraries on the Web. Bulletin of the Medical Library Association 86(4):541-4, 1998.

4. Yan TW, Jacobsen M, Garcia-Molina H, Dayal U. From user access patterns to dynamic hypertext linking. Fifth International World Wide Web Conference,1996, http:// www5conf.inria.fr/ fich_html/ papers/P8/Overview.html.

5. D'Allessandro MP, D'Alessandro DM, Galvin JR, Erkonen WE. Evaluating overall usage of a digital health sciences library. Bulletin of the Medical Library Association 86(4):602-9, 1998.

6. Lawrence S, Giles CL. Context and page analysis for improved Web search. IEEE Internet Computing, July-August:38-46, 1998.

7. Cooper MD. Design considerations in instrumenting and monitoring Web-based information retrieval systems. Journal of the American Society for Information Science 49(10):903-919, 1998.

8. Jansen BJ, Spink A, Saracevic T. Failure analysis in query construction: Data and analysis from a large sample of Web queries. Proceedings of the Third ACM Conference on Digital Libraries, 289-90, 1998.

9. UMLS Knowledge Sources, 10th ed., Bethesda, MD, National Library of Medicine, 1999.

10. McCray AT, Razi AM, Bangalore AK, Browne AC, Stavri PZ. The UMLS Knowledge Source Server: A versatile Internet-based research tool. Proceedings of the 1996 AMIA Fall Symposium, 164-8, 1996.

11. Schwartz C. Web search engines. Journal of the American Society for Information Science, 49(11):973-982, 1998.

12. Schuyler PL, McCray AT, Schoolman HM. A test collection for experimentation in bibliographic retrieval. In: Proceedings of the Sixth Conference on Medical Informatics: North Holland, 910-12, 1989.

APPENDIX

20 most frequent queries in the three month data set

392 visible human	210 aids
384 index medicus	210 lupus
376 diabetes	208 brain
302 fibromyalgia	203 breast cancer
297 cancer	195 creatine
262 heart	188 thyroid
242 multiple sclerosis	184 viagra
239 anatomy	180 medline
226 hepatitis c	173 hypertension
215 asthma	163 nursing